

Gap-weighted subsequences for automatic cognate identification and phylogenetic inference

Taraka Rama
Språkbanken
University of Göteborg

Abstract

In this paper, we describe the problem of cognate identification and its relation to phylogenetic inference. We introduce subsequence based features for discriminating cognates from non-cognates. We show that subsequence based features perform better than the state-of-the-art string similarity measures for the purpose of cognate identification. We use the cognate judgments for the purpose of phylogenetic inference and observe that these classifiers infer a tree which is close to the gold standard tree. The contribution of this paper is the use of subsequence features for cognate identification and to employ the cognate judgments for phylogenetic inference.

1 Introduction

Historical linguistics, the oldest branch of modern linguistics, studies how languages change and attempts to infer the genetic relationship between languages with suspected relationship. In this context, genetic relationship means that two languages are solely similar due to their descent from a common ancestor and not due to structural similarity. Identification of cognates is a very important step prior to the positing of any genetic relationship between two languages.

Cognates are words across languages whose origin can be traced back to a common ancestor. For example, English *night* ~ German *Nacht* ‘night’ and English *hound* and German *Hund* ‘dog’ are cognates whose origin can be traced back to a common ancestor. In historical linguistics, cognates are identified through the application of the comparative method (Rankin, 2003). Sometimes, cognates are not revealingly similar but have changed substantially over time such that

they do not share form similarity. An example of such a cognate pair is the English *wheel* and Sanskrit *chakra* ‘wheel’, which can be traced back to Proto-Indo-European (PIE) $*k^wek^welo$.

When historical linguists work with the comparative method, they compare basic vocabulary, phonological correspondences, grammatical forms, and morphological paradigms to establish relationship between languages suspected of common descent. However, performing a large scale automatic grammatical correspondence analysis presupposes that we have well-defined morphological analyzers for ancient, extinct, and underdocumented languages.

Basic vocabulary lists such as the ones devised by Morris Swadesh (Swadesh, 1952), provide a suitable testing ground for applying machine learning algorithms to automatically identify cognates. Some standardized word lists come with cognate information and, subsequently, is used to infer the relationship between languages under purview (Dyen et al., 1992). In the related field of computational biology, the term *phylogenetic inference* is in vogue to signify computational methods which infer relationship between biological species (Felsenstein, 2004). The same term has come to refer to the identification of genetic relationships between languages also (McMahon and McMahon, 2005).

Swadesh (1952) developed lexicostatistics as a technique to infer relationships between languages. In this effort, Swadesh posited a list of basic vocabulary items, ranging from sizes 100–200 that are supposed to be universal, culture-free, and resistant to replacement over time. In positing these word lists, Swadesh intended to develop a concept list where the translational equivalents for each language would be provided by language experts. In the next step, the cognacy status between a pair of words is determined through the application of the comparative method. Finally,

the similarity of a language pair is defined as the total number of shared cognate word pairs divided by the total number of word pairs. The pair-wise distance matrix computed from this step can then be supplied to a clustering algorithm such as UP-GMA (Sokal and Michener, 1958)¹ to infer a tree between the languages.

Thus, the tasks of establishing relationship between languages as well as the identification of cognates are closely related tasks where the output of the latter serves as an input to the former. Automatic cognate identification, as defined in computational linguistics literature, refers to the application of string similarity or phonetic similarity algorithms either independently, or in tandem with machine learning algorithms for determining if a given word pair is cognate or not (Inkpen et al., 2005).²

The approaches developed by Kondrak and Sherif (2006) and Inkpen et al. (2005) supply different string distances between a pair of words as features to a linear classifier. Usually, a linear classifier such as Support Vector Machine (SVM) is trained with labeled positive (“cognates”) and negative (“non-cognates”) examples and tested on a held-out dataset. Cognate information has been applied to the tasks of sentence alignment (Simard et al., 1993) and statistical machine translation (Kondrak et al., 2003).

In this paper, we use subsequence based features for automatic cognate identification as well as phylogenetic inference. We show that subsequence based features outperform word similarity measures for the task of cognate identification. We motivate the use of subsequence based features in terms of linguistic examples and then proceed to formulate the subsequence based features that can be derived from string kernels (Shawe-Taylor and Cristianini, 2004) developed for text categorization task (Lodhi et al., 2002). In IR literature, string subsequences go under the name of skipgrams (Järvelin et al., 2007).

¹Also known as average-linking clustering in NLP (Manning and Schütze, 1999).

²In NLP, even borrowed words (*loanwords*) which usually have strong semantic as well as form similarity are referred to as cognates. In contrast, historical linguistics makes a stark distinction between loanwords and cognates. An example of a loanword is English *beef* from Norman French. *correlates* (McMahon and McMahon, 2005) include cognates and borrowings. In NLP, the words descending from an ancestral language are referred to as ‘genetic cognates’ (Kondrak, 2005). In this paper, we use cognates to refer to those words whose origin can be traced back to a common ancestor.

The rest of the paper is structured as followed. In section 2, we define the two problems of automated cognate identification and phylogenetic inference. We describe related work in section 3. Section 4 describes subsequence features, experimental setup, dataset, evaluation measures, and results. In section 5, we describe our phylogenetic experiment setup and the evaluation measure for the inferred tree. We discuss the results of our experiments as we present them. Finally, we conclude and provide pointers to future direction in section 6.

2 Two problems

In this paper, we work with identifying cognates in Swadesh lists for the Indo-European family. The Swadesh lists – of length 200 – for 84 Indo-European languages were compiled by Dyen et al. (1992). As mentioned before, the Swadesh lists contain the lexical realization for a concept and its cognate class. A cognate class is a function mapping a set of multiple items belonging to different languages to a unique cognate class number (CCN). Hence, for each concept, positive training instances consist of pairs of words belonging to different languages that share a CCN. If the words in the pair do not share a CCN number, then the word pair is labeled as a negative instance. We intend to explore the efficacy of subsequence features to the following problems:

1. In a scenario where there are few positive examples and a very large number of negative examples, how well do subsequence features perform over string similarity measures in the task of cognate identification?
2. In many families, the information about cognacy judgments is partially available. In such a case, how well can a classifier trained on partial data be used to identify cognates in the remaining languages? Can the classifier generalize over the language family? Can the cognate judgments inferred from the previous step be used to infer a phylogenetic tree?

3 Related work

Ellison and Kirby (2006) use scaled edit distance (normalized by average length) to measure the intra-lexical divergence in a language. This step yields a language-internal probability distribution. They then apply the KL-divergence measure to calculate the distance between a language pair.

This step is repeated for all the 42×83 language pairs from Dyen et al.’s IE database to yield a distance matrix. The distance matrix is then used to infer a tree for the IE language. Unfortunately, they perform a qualitative evaluation of the inferred tree and do not compare the tree to the standard tree inferred by experts of the language family. The authors mention string kernels but do not pursue this line of research further.

Bouchard-Côté et al. (2013) employ a graphical model to reconstruct the proto-word forms from the synchronic word-forms for the Austronesian language family. They compare their automated reconstructions with the ones reconstructed by historical linguists and find that their model beats an edit-distance baseline. However, their model has a strict requirement that the tree structure between the languages under study has to be known beforehand.

Greenhill (2011) argues against the use of vanilla edit distance for cognate identification and language distance computation. However, a recent paper by Hauer and Kondrak (2011) shows that a combination of edit distance and other string similarity measures, supplied as features to a SVM classifier, will boost the cognate identification accuracy.

Hauer and Kondrak (2011)³ supply different string similarity scores as features to a SVM classifier for determining if a given word pair is a cognate or not. The authors also employ an additional binary language-pair feature – that is used to weigh the language distance – and find that the additional feature assists the task of semantic clustering. In this task, the cognacy judgments given by a linear classifier is used to flat cluster the lexical items belonging to a single concept. The clustering quality is evaluated against the gold standard cognacy judgments. Unfortunately, the experiments of these scholars cannot be replicated since the partitioning details of their training and test datasets is not available.

In our experiments, we use edit distance as the sole feature for a baseline classifier. We also compare our results with the results of the classifiers trained from HK-based features.

4 Cognate identification

The vanilla edit distance measure counts the minimum number of insertions, deletions, and substi-

tutions required to transform a word into another word. Identical words have 0 edit distance. For example, the edit distance between two cognates English *hound* and German *hund* is 1. Similarly, the edit distance between Swedish *i* and Russian *v* ‘in’, which are cognates, is 1. The edit distance treats both of the cognates at the same level and does not reflect the amount of change which has occurred in the Swedish and Russian words from the PIE word.

Another string similarity measure such as Dice⁴ estimates word similarity as the ratio between the number of common bigrams to estimate the similarity between two words. The similarity between Lusatian *dolhi* and Czech *dluhe* ‘long’ is 0 since they do not share any common bigrams and the edit distance between the two strings is 3. Although the two words share all the consonants, the Dice score is 0 due to the intervening vowels.

Another string similarity measure, Longest Common Subsequence (LCS) measures the length of the longest common subsequence between the two words. The LCS is 4 (*hund*), 0, and 3 (*dlh*) for the above examples. One can parade a number of examples which are problematical for the simple-minded string similarity measures. Alternatively, string kernels in machine learning research offer a way to exploit the similarities between two words without any restrictions on the length and character similarity.

4.1 Subsequence features

Edit distance in its rawest form aligns two strings based on the minimum number of edit operations. Edit distance neither makes any distinction between aligning vowels to consonants nor does it account for the similarity between two sounds (e.g., /p/ and /b/). Multiple approaches have been proposed to alleviate these shortcomings. Wieling et al. (2009) propose a Vowel-Consonant-constrained edit distance, based on PMI (pair-wise mutual information), for the purpose of extracting matching sounds between two words.⁵ They apply their method to dialect data and find that their method identifies the traditional dialectal boundaries. In extension, Jäger (2013) used a PMI-based edit distance on a training dataset to compute the distance between phonetic symbols. The sym-

³Henceforth, referred to as HK.

⁴In general, Dice between two sets is defined as the ratio of number of shared elements to the total number of elements in both the sets.

⁵Vowels do not align with consonants.(Prokić, 2010)

bol similarity matrix was used to compute pair-wise language distances. The pair-wise language distances were then compared to the gold standard classification. They find that PMI-based edit distance outperforms edit-distance based language distances.

Turchin et al. (2010) employ matching consonant classes to determine the similarity between two words. These approaches require explicit formalization of linguistic constraints depending on the languages under consideration. In fact, vowel quality is known to vary across time. If we drop the vowels in the Czech-Lusatian word pair, then the words are identical. In another study, List (2012) uses a permutation based method to learn the similarity between sounds and employs the technique to cluster identified cognates for a concept. A SVM classifier learns the weight for a subsequence feature and combines the learned weights of the features without any human intervention.

Subsequences of length greater than 1 also take context into account. Subsequences as formulated below weigh the similarity between two words based on the number of dropped characters and combine vowels and consonants seamlessly. Having motivated why subsequences seems to be a good idea, we formulate subsequences below.

We follow the notation given in Shawe-Taylor and Cristianini (2004) to formulate our representation of a word (string). Given a string s , the subsequence vector $\Phi(s)$ is defined as follows. The string s can be decomposed as $s_1, \dots, s_{|s|}$ where $|s|$ denotes the length of the string. Let \vec{l} denote a sequence of indices $(i_1, \dots, i_{|\vec{l}|})$ where, $1 \leq i_1 < \dots < i_{|\vec{l}|} \leq |s|$. Then, a subsequence u is a sequence of characters $s[\vec{l}]$. Note that a subsequence can occur multiple times in a string. Then, the weight of u , $\phi(u)$ is defined as $\sum_{\vec{l}: u=s[\vec{l}]} \lambda^{l(\vec{l})}$ where, $l(\vec{l}) = i_{|\vec{l}|} - i_1 + 1$ and $\lambda \in [0, 1]$ is a decay factor. The subsequence vector $\Phi(s)$ is $(\phi_{u_1} \dots \phi_{u_{|\Sigma^*|}})$ where, $\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$ is the set of all strings from an alphabet Σ . In our experiments, we fix the value of λ at 0.5.

The λ factor is exponential and penalizes u over long gaps in a string. Due to the above formulation, the frequency of a subsequence u is also taken into account. In our experiments, we observed that a few thousand word pairs did not have a single character in common. In such a scenario, we default to class-based subsequence features by

mapping a Σ in u to its Consonant/Vowel class – $\Sigma \mapsto \{C, V\}$. As a preliminary step, we map each string s into its C, V sequence s_{cv} and then compute the subsequence weights.⁶

A combined subsequence vector $\Phi(s + s_{cv})$ is further normalized by its norm, $\|\Phi(s + s_{cv})\|$, to transform into a unit vector. The common subsequence vector $\Phi(s_1, s_2)$ is composed of all the common subsequences between s_1, s_2 . The weight of a common subsequence is $\phi_u^{s_1} + \phi_u^{s_2}$.

Moschitti et al. (2012) list the features of the above weighting scheme.

- Longer subsequences receive lower weights.
- Characters can be omitted (called gaps).
- The exponent of λ penalizes recurring subsequences with longer gaps.

For a string of length m and a pre-determined subsequence length p , the computational complexity is in the order of $\mathcal{O}(mp)$.

On a linguistic note, gaps are consistent with the prevalent sound changes such as sound loss, sound gain, and assimilation⁷, processes which alter word forms in an ancestral language causing the daughter languages to have different surface forms. The λ factor weighs the number of gaps found in a subsequence. For instance, the Sardinian word form for ‘fish’ *pissi* has the subsequence *ps* occurring twice but with different weights: λ^3, λ^4 .

The combined feature vector, for a word pair, is used to train a SVM classifier. In our experiments, we use the LIBLINEAR package (Fan et al., 2008) to solve the primal problem with L_2 -regularization and L_2 -loss. The next subsection describes the makeup of the dataset. We use the default parameters since we did not observe any difference in our development experiments.

4.2 Dataset

We used the publicly available Indo-European dataset (Dyen et al., 1992) for our experiments. The dataset has 16,520 lexical items for 200 concepts and 84 language varieties. Each word form is assigned to a unique CCN. A concept can have multiple word forms. In such a case, we randomly pick one word and discard the rest of the forms. There are more than 200 identical non-cognate pairs in the dataset.

⁶ $V = \{a, e, i, o, u, y\}, C = \Sigma \setminus V$

⁷A sound can assimilate to a neighboring sound. Sanskrit *agni* > Prakrit *aggi* ‘fire’. Compare the Latin form *ignis*.

For the first experiment, we extracted all word pairs for a concept and assigned a positive label if the word pair has an identical CCN; a negative label, if the word pair has different CCNs. We extracted a total of 674,192 word pairs out of which 158,787 are cognates.

The word length is an important parameter in our experiments since it gives an index of how far the value of subsequence length, p , should be tested. We found that the average word length is about 4.79 and the median is 5. There are about 928 words which have a word length greater than 7. Hence, we tested the effect of p from 1 to 7. We report the results for different values of p .

Subfamily	# of languages
Germanic	14
Indo-Iranian	18
Romance	14
Slavic	13
Others	25

Table 1: Number of languages in each subfamily.

The second experiment involves cognate identification as a step towards phylogenetic inference. In this experiment, we split the 84 languages into training and test sets based on their membership in subfamilies. The IE dataset has 84 languages belonging to 8 different subfamilies. Out of these, Germanic, Indo-Iranian, Romance, and Slavic have more than 10 languages (cf. table 1). The rest of the languages are distributed across the Celtic, Baltic, Armenian, and Albanian groups.

	positive (+ve)	negative (-ve)
training	38,722	135,658
test	39,432	119,389

Table 2: Number of positive and negative examples in the training and test datasets.

We merged all the groups with less than 10 languages into a single group of 25 languages, “Others”. Then, we randomly split each subfamily into a training and testing dataset of roughly equal size. Subsequently, we merged the subfamilies’ training datasets into a single training dataset. We followed the same merging procedure with the test datasets to create a single test dataset for the whole language family. Finally, we extracted the subsequence feature vectors for each labeled word pair. The details of dataset is given in table 2. The idea behind this setup is explained in question 2 under section 2.

4.3 Evaluation measures

In this section, we describe the different measures for evaluating the results of our experiments. In our first experiment, the performance of the linear classifier was evaluated using five-fold cross-validation accuracy. The accuracy measure is defined as below:

$$N = TN + TP + FN + FP \quad (1a)$$

$$ACC = \frac{TP + TN}{N} \quad (1b)$$

where, TP: number of true positives, TN: true negatives, FP: false positives, FN: false negatives, and N shows the total number of test instances.

In the second experiment of cognate identification, we use Matthews Correlation coefficient (MCC) and Average Precision (AP) for evaluating the performance of our classifier. MCC (Matthews, 1975) is a comprehensive evaluation measure which takes TP, TN, FP, and FN into account when computing the agreement between the predicted binary vector, \hat{y} and the gold standard binary vector, y . The calculation of MCC is not straightforward and is given in equations 2a–2c. $MCC \in [-1, +1]$ where a score of -1 suggests perfect disagreement whereas $+1$ suggests perfect agreement. MCC is used when there is a difference in the size of the classes in the test dataset. In our case, the number of negative examples are thrice the number of positive examples. MCC is a special case of Pearson’s r , which measures the agreement between two binary vectors.

$$S = \frac{TP + TN}{N} \quad (2a)$$

$$P = \frac{TP + FP}{N} \quad (2b)$$

$$MCC = \frac{TP/N - S \times P}{\sqrt{PS(1-P)(1-S)}} \quad (2c)$$

The classification score given by a linear classifier for a test instance is transformed into a probability score through sigmoid function. Thus a test instance is labeled as positive if it has a probability score > 0.5 , else is classified as negative. In a classic information retrieval task setting, one would look at the precision, $p(r)$, plotted as a function of recall $r \in [0, 1]$, to observe the performance of the classifier for various classifier thresholds. Hence, we report the average precision (AP) score, defined as $\int_0^1 p(r)dr$, for each experimental setting.

It is worth noting that Kondrak (2009) employs 11-point interpolated average precision for evaluating different similarity algorithms on a small test set consisting of five languages. In our experiments, we use a larger test set of 41 languages. The AP score also measures the robustness of a classifier against different thresholds. If a classifier ranks low at AP but evaluates well for other measures, it suggests that the classifier is not robust to the shifting probability thresholds.

We use the three evaluation measures to check if the classifiers perform well on the task of detecting TPs and TNs. Ideally, a cognate identifier system should perform well on both positive and negative examples. The difference between MCC and AP is that MCC evaluates the performance of a classifier on both positive and negative examples for a fixed threshold.

4.4 Results

In this subsection, we describe the results of our experiments on cognate identification using subsequence features. In the first experiment, we perform a five-fold cross-validation on the labeled positive and negative examples. Then we move to report our results on the combined feature vectors comprising subsequence features and word pair similarity features. The following word pair similarity features from Hauer and Kondrak (2011) are used in our experiments:

- Edit distance
- Length of longest common prefix
- Number of common bigrams
- Lengths of individual words
- Absolute difference between the lengths of the words

which is referred to as HK in all the tables.

4.4.1 Cross-validation experiments

The main aim of this experiment is to determine if subsequence features work at least as well as HK features on a dataset split into five folds. The accuracies presented in table 3 show that subsequence features work better than HK for all values of p . The highest accuracy is at $p = 7$. Even the subsequences of length 2 outperform the HK-based classifier and baseline classifier. In the rest of the paper, we do not use the baseline classifier but compare our results against the HK classifier.

Encouraged by this positive result, we proceeded to test if the combination of HK features and subsequence features improve the cross-

Features	ACC
Baseline	77.4239
HK	82.2976
1	81.8971
2	83.3375
3	83.4291
4	83.5284
5	83.5393
6	83.5382
7	83.5682

Table 3: Five-fold cross-validation accuracy for various lengths of p .

validation accuracies. The results of this experiment is shown in table 4. In this experiment, the highest result is for $p = 4$. We report the results for $p > 1$ since, the $p = 1$ classifier performs worse than HK-based classifier. These re-

Features	ACC
HK+2	84.0117
HK+3	84.044
HK+4	84.047
HK+5	84.0427
HK+6	84.0448
HK+7	84.027

Table 4: Five-fold cross-validation accuracy for a combination of HK features and various subsequence lengths.

sults show the superiority of subsequence features over word similarity features. Now, we move on to test the performance of subsequence features in a real-world scenario.

4.4.2 Subfamily experiments

The datasets used in these experiments are intended to imitate the real world situation where there are gaps in knowledge regarding the cognate status of word pairs. In NLP, the default ratio between training and test datasets is 4:1 or 3:1. In comparative linguistics, the amount of available labeled data would be much less. Hence, a 50-50 random split of the language groups tests the efficiency of subsequence features for cognate identification and phylogenetic inference. We perform two sets of experiments with the randomly split language groups.

The first set of experiments consist of testing the performance of subsequence based features against HK features on multiple aspects. The results of this experiment is given in table 5. The results suggest that the subsequence features perform consistently over $p \in [2, 7]$. The $p = 3$ based linear classifier performs the best across

all the evaluation measures. All the subsequence based features agree on AP score and perform better than HK classifier. The subsequence-based features outperform at MCC and ACC evaluation measures.

Features	ACC	MCC	AP
HK	81.2034	0.4269	0.6565
2	81.9048	0.4542	0.6662
3	82.0968	0.4618	0.6674
4	81.9451	0.4558	0.6664
5	81.9533	0.4561	0.6665
6	81.9117	0.4546	0.6663
7	81.9281	0.4552	0.6664

Table 5: Performance of subsequence features on subfamily test set.

We tested if the results of $p = 3$ classifier is better than the HK classifier using a *paired t-test*. A classifier’s agreement/disagreement (1/0) with gold standard classification is encoded as a binary vector. Then, a paired t-test is used to determine if there is a statistically significant difference between the two classifiers. The difference between HK and $p = 3$ is significant at the 0.001 level. Now we move to the results of the combination experiment.

In this experiment, we use the same training and testing set but use the feature combination explored in the cross-validation experiments. In these experiments, the HK+2-based classifier won across all the evaluation measures. The combination classifiers perform similarly on all the evaluation measures. We ranked HK+2 classifier for the reason that the classifier has lesser number of parameters and can be computed in lesser time than the rest of the classifiers.

Features	ACC	MCC	AP
HK+2	82.8121	0.4877	0.7017
HK+3	82.8039	0.4872	0.7015
HK+4	82.7655	0.4857	0.7012
HK+5	82.7674	0.4858	0.7012
HK+6	82.7649	0.4857	0.7012
HK+7	82.7655	0.4857	0.7012

Table 6: Performance of combination of subsequence and HK features on subfamily test set.

A paired t-test shows that the difference between HK and HK+2 classifier’s predictions are significant at the 0.001 level. Also, the difference between the HK+2 and $p = 3$ classifiers is significant at the 0.001 level. We conclude by observing that subsequence-based classifiers perform better than a HK-based classifier.

Now, we proceed to do an error analysis and then attempt to use our cognate judgments for the purpose of phylogenetic inference described in the next section.

4.5 Error analysis

In this section, we examine the misclassified word pairs. Our hypothesis is that majority of FPs are correlates and FNs are those items which are quite dissimilar. The gold standard cognate classification of a word pair is binary in nature and cannot be used to measure the exact form similarity of a word pair. In lieu, we use length normalized edit distance (LDN) to measure the difference. To test our hypothesis about FNs and FPs, we correlated the classifier scores of word pairs in each error class and classifier with the corresponding length normalized edit distance scores. We expect a negative correlation between the classifier scores and LDN scores since the former are similarity scores. In fact, the correlations are negative as in table 7.

Classifier	FP	FN
$p = 3$	-0.29 (0.56)	-0.42 (0.38)
HK+2	-0.55 (0.55)	-0.48 (0.38)

Table 7: Correlation between probability scores and LDNs. The average of a classifier’s probabilities is shown in (...).

5 Phylogenetic inference

We describe a popular tree inference algorithm known as Neighbor-Joining (NJ) algorithm (Saitou and Nei, 1987). Then, we describe our gold standard tree and Generalized Quartet distance (GQD) for measuring the distance between the inferred tree and the gold standard tree.

5.1 Tree inference

The cognate judgments returned by the linear classifier can be used to compute the distance between a distance matrix, D , containing the distances between all the language pairs in the test set. We can define binary and similarity-based distance matrices from the classifier judgments. The binary distance d_{ij}^b between languages i, j is defined as $1 - \frac{|\{k|\hat{y}_k=1\}|}{n_{ij}}$, where n_{ij} is the total number of word pairs between i, j . As mentioned earlier, the sigmoid function maps the linear score of a classifier into $p(\hat{y}_k) \in [0, 1]$. $p(\hat{y}_k)$ can be used to define the classifier distance d_{ij}^s as $1 - \frac{\sum_k p(\hat{y}_k)}{n_{ij}}$.

The matrix D is then supplied as an input to the NJ algorithm⁸ to infer a tree between the languages. The test set has 41 languages and there are $\approx 10^{19}$ possible unrooted trees for 40 languages. The problem of exact tree search is a computationally hard problem and there exist heuristic techniques to reduce the searchable tree space.

NJ algorithm is a clustering algorithm which has a complexity of $\mathcal{O}(l^3)$, l is the number of languages, and is shown to converge quickly for biological datasets consisting of thousands of species. NJ has been widely tested over both real and simulated datasets and was reported to be statistically consistent over different test conditions.

5.2 Gold standard tree

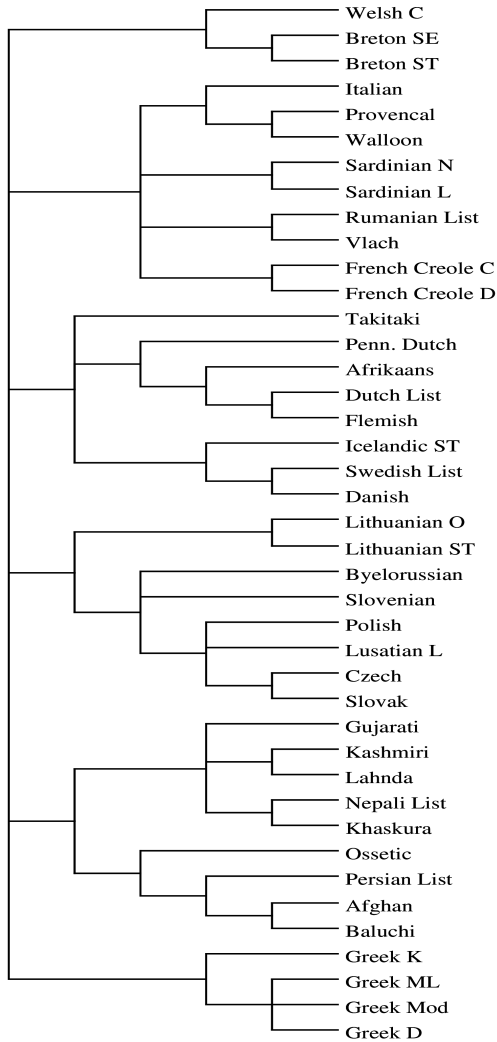


Figure 1: The gold standard classification of the test set.

⁸Available on <http://splitstree.org/>

The gold standard classification is shown in figure 1. Although, Dyen et al., provide a classification for the languages in the test set, the classification wrongly places the languages in the tree. Hence, we extract the relevant languages from the expert classification given in Nordhoff and Hammarström (2012). The highest level of the tree is polytomous or shows non-binary branching. The nature of highest level branching is still an open question in Indo-European historical linguistics. Hence, our gold standard tree also shows the gaps in the scholarship. In fact, the tree evaluation metric that will be introduced in the next subsection attempts to alleviate this issue.

5.3 Tree distance measure

In evolutionary biology, tree distance measures are used to measure the accuracy of a tree inference algorithm. Quartet distance is the state-of-the-art tree distance measure used to compute the distance between two trees. Quartet distance is defined as the number of different quartets between the trees. A quartet is a subtree with four leaves and there are $\binom{l}{4}$ quartets in a tree with l leaves.

A quartet is *resolved* if there exists an internal node that separates a pair of leaves. For example, the quartet consisting of Swedish, Danish, Icelandic, and Dutch is resolved since Swedish and Danish are separated from Icelandic and Dutch through an internal node. Such a quartet is known as a *butterfly* quartet. A star quartet is complementary to a butterfly quartet since all the languages in a star quartet are connected to a central node. The top node in the figure 1 is an example of a star quartet.

The quartet distance (QD) between two trees, T_1, T_2 is defined as:

$$\frac{q(T_1) + q(T_2) - 2s(T_1, T_2) - d(T_1, T_2)}{\binom{l}{4}} \quad (3)$$

where $q(T)$ is the number of butterflies in T , $s(T_1, T_2)$ is the number of shared butterflies between T_1, T_2 , and $d(T_1, T_2)$ is the number of different butterflies between T_1, T_2 .

Christiansen et al. (2006) developed a fast algorithm for computing the quartet distance between trees having thousands of leaves. The QD formula in equation 3 counts the number of resolved quartets in the inferred tree as errors. The inferred binary tree T_i should not be penalized for the unresolvedness in the gold standard tree T_g . Pompei et al. (2011) defined a new measure known as GQD

to negate the effect of star quartets in T_g . GQD is defined as $d(T_i, T_g)/q(T_g)$. We use both QD and GQD to evaluate the quality of the inferred trees.

5.4 Tree inference results

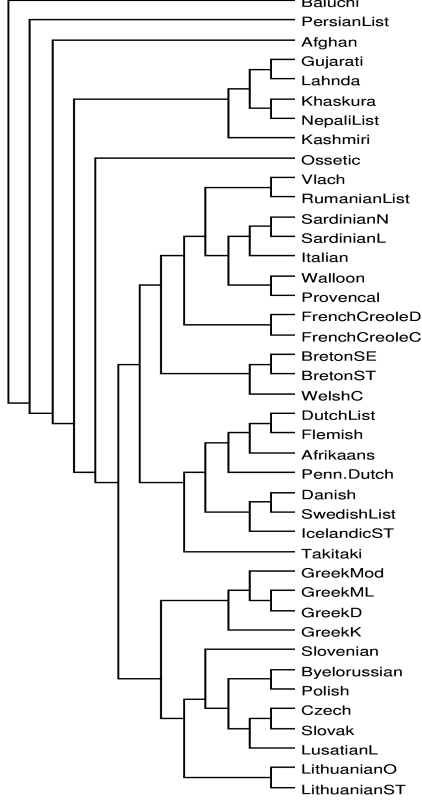


Figure 2: The best tree on the test set based on HK (binary) classifier.

Classifier	QD	GQD
Oracle	0.29766	0.005648
HK (wts.)	0.30236	0.012303
HK (binary)	0.301669	0.011324
$p = 3$ (wts.)	0.303782	0.014316
$p = 3$ (binary)	0.304246	0.014973
HK+2 (wts.)	0.30236	0.012303
HK+2 (binary)	<i>0.30234</i>	<i>0.012275</i>

Table 8: QD and GQD for the top performing models in subfamily experiment.

In this experiment, we compute distance matrices from the predictions of the top classifiers in section 4.4.2. The winning classifiers are subsequence classifiers: $p = 3$ and HK+2. We compare the winning classifiers against HK classifier. The output of each classifier is used to compute the binary and weighted (shown in table 8 as “(wts.)”) distance matrices based on equations de-

finied above. In order to see the effect of tree inference algorithm, we also report the difference between the tree inferred from gold standard cognate judgments (*Oracle tree*) and the gold standard tree given by expert historical linguists.

Each phylogenetic tree is compared to the gold standard tree (cf. figure 1) using the tree distance measures, QD and GQD. All the classifiers give similar results in this experiment. The results suggest that the choice of binary vs. weighted cognacy judgments do not make a significant difference in the quality of the inferred trees. The results for HK-based classifier are shown in bold-face since, it gives the best result and is also the simplest of all the classifiers in terms of model complexity. The oracle tree also differs from the expert classification.

The main difference between the HK (binary) tree and the next best tree (italicized results in table 8) is the placement of Takitaki language. Both trees misplace Ossetic as an outlier in the Indo-Iranian branch whereas, it should have been placed together with Iranian branch. The HK+2 (binary) tree places Byelorussian correctly whereas HK (binary) tree misplaces it. Italian’s position is correctly determined in HK+2 (binary) tree whereas HK tree misplaces it. Overall, the difference between the top-two trees is not large.

6 Conclusion and future work

In this paper, we introduced subsequences and tested their efficacy for cognate identification and phylogenetic inference in a scenario where there is incomplete knowledge about a language family. We showed that subsequences perform significantly better than simple word similarity based classifiers for cognate identification. We evaluated the performance of the classifiers at the task of phylogenetic inference and found that there is no significant difference between the various classifiers.

As a future work, we intend to employ fuzzy subsequence matching for building the feature vectors for a word pair using a phonetic similarity measure. We also intend to integrate articulatory features of sounds into our experiments. We plan to test our features on Austronesian vocabulary lists (Greenhill et al., 2008). Further, we plan to test the subsequence features for automated classification of thousands of languages available in ASJP database (Wichmann et al., 2010).

Acknowledgments

I warmly thank Richard Johansson, Johann-Mattis List, and Søren Wichmann for all the comments which made the draft better. The paper was originally submitted to EMNLP 2014 but was rejected. I benefitted substantially from the comments made by all the three people.

References

- [Bouchard-Côté et al.2013] Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.
- [Christiansen et al.2006] Chris Christiansen, Thomas Mailund, Christian NS Pedersen, Martin Randers, and Martin Stig Stissing. 2006. Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology*, 1(1).
- [Dyen et al.1992] Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indo-European classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5):1–132.
- [Ellison and Kirby2006] T. Mark Ellison and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 273–280, Sydney, Australia, July. Association for Computational Linguistics.
- [Fan et al.2008] Rong-En Fan, Kai-Wei Chang, Chong-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.
- [Felsenstein2004] Joseph Felsenstein. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.
- [Greenhill et al.2008] Simon J. Greenhill, Robert Blust, and Russell D. Gray. 2008. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics Online*, 4:271–283.
- [Greenhill2011] Simon J. Greenhill. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4):689–698.
- [Hauer and Kondrak2011] Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 865–873, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.
- [Inkpen et al.2005] Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257.
- [Jäger2013] Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.
- [Järvelin et al.2007] Anni Järvelin, Antti Järvelin, and Kalervo Järvelin. 2007. s-grams: Defining generalized n-grams for information retrieval. *Information Processing & Management*, 43(4):1005–1019.
- [Kondrak and Sherif2006] Grzegorz Kondrak and Tarek Sherif. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of ACL Workshop on Linguistic Distances*, pages 43–50. Association for Computational Linguistics.
- [Kondrak et al.2003] Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003—short papers*, volume 2, pages 46–48. Association for Computational Linguistics.
- [Kondrak2005] Grzegorz Kondrak. 2005. Cognates and word alignment in bitexts. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 305–312.
- [Kondrak2009] Grzegorz Kondrak. 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement automatique des langues et langues anciennes*, 50(2):201–235, October.
- [List2012] Johann-Mattis List. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France, April. Association for Computational Linguistics.
- [Lodhi et al.2002] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444.
- [Manning and Schütze1999] Christopher D Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

- [Matthews1975] Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.
- [McMahon and McMahon2005] April McMahon and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford University Press, Oxford.
- [Moschitti et al.2012] Alessandro Moschitti, Qi Ju, and Richard Johansson. 2012. Modeling topic dependencies in hierarchical text categorization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 759–767. Association for Computational Linguistics.
- [Nordhoff and Hammarström2012] Sebastian Nordhoff and Harald Hammarström. 2012. Glottolog/Langdoc: Increasing the visibility of grey literature for low-density languages. In *Language Resources and Evaluation Conference*, pages 3289–3294.
- [Pompei et al.2011] Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS ONE*, 6(6):e20109.
- [Prokić2010] Jelena Prokić. 2010. *Families and Resemblances*. Ph.D. thesis, Rijksuniversiteit Groningen.
- [Rankin2003] Robert L. Rankin. 2003. The comparative method. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 199–212. Wiley Online Library.
- [Saitou and Nei1987] Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.
- [Shawe-Taylor and Cristianini2004] John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge university press.
- [Simard et al.1993] Michel Simard, George F. Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing*, volume 2, pages 1071–1082. IBM Press.
- [Sokal and Michener1958] Robert R Sokal and Charles D Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.
- [Swadesh1952] Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4):452–463.
- [Turchin et al.2010] Peter Turchin, Ilja Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, 3:117–126.
- [Wichmann et al.2010] Søren Wichmann, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, 389:3632–3639.
- [Wieling et al.2009] Martijn Wieling, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 26–34. Association for Computational Linguistics.